Statistical analysis of high-dimensional biomedical data: A gentle introduction

Federico Ambrogi[1], Jörg Rahnenfuehrer[2], Riccardo De Bin[3], Lisa McShane[4], on behalf of the TG9 – High-Dimensional Data of the STRATOS initiative

[1]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

[2]Department of Statistics, TU Dortmund University, Dortmund, Germany

[3]Department of Mathematics, University of Oslo, Oslo, Norway

[4]Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Outline of the talk

- TG9 Overview
- Updates
  - Members
  - Overview paper
- Current work
  - Sample size calculation in HDD
  - Influence and choice of the tuning parameters
  - Use of plasmode data for simulations in HDD

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE
E DI COMUNITÀ

TG9 Overview

- The goal of the High-dimensional Data (HDD) Topic Group of the STRATOS initiative (TG9) is to provide guidance amid the jungle of opportunities and pitfalls inherent in the analysis of high-dimensional biological and medical data.
- Methods for analysis of HDD are rapidly changing, and researchers across different fields, including biostatistics, bioinformatics, and bioengineering, contribute to their development.
- Advances in statistical methodology and machine learning methods have contributed to improved approaches for data mining, statistical inference, and prediction in HDD settings;
- Adoption of these method has sometimes gotten ahead of understanding of their proper application.
- The mission of TG9 includes identification of fundamental principles for analysis of HDD, explanation of available methods, and development of broadly accessible guidance on best practices in this complex and changing landscape.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Updates:

members

TG9 current members:

Federico Ambrogi (Italy)
Axel Benner (Germany)
Harald Binder (Germany)
A.-L. Boulesteix (Germany)
Riccardo De Bin (Norway)
Kevin Dobbin (USA)
Ilaria Gardin (Italy)
Roman Hornung (Germany)
Lara Lusa (Slovenia)
Lisa McShane (USA)
Stefan Michiels (France)
E. Migliavacca (Switzerland)
Jörg Rahnenführer (Germany)
Willi Sauerbrei (Germany)

- 14 members;
- 3 new members (in italics);
- 3 co-chairs (in bold).

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Updates:

Overview paper

In the last meetings, we used Molière's words

"Trees that are slow to grow bear the best fruit."

to describe the state of our 11-author, ∼40,000-word, ∼60-page manuscript ...

The "fruit" is finally ripe and the paper has been accepted in BMC Medicine.

- Title: *Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges*;
- this review provides a solid statistical foundation for researchers, including statisticians and non-statisticians, who are new to research with HDD or simply want to better evaluate and understand the results of HDD analyses.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Updates:

Overview paper

Table of contents:

1. Background
   1.1. Study Design
2. Methods
   2.1. IDA: Initial data analysis and preprocessing
   2.2. EDA: Exploratory data analysis
   2.3. TEST: Identification of informative variables and multiple testing
   2.4. PRED: Prediction
3. Discussion
4. Conclusions

Each section contains analytical goals, common approaches, examples and gaps, related to the specific stage of the HDD analysis.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Overview paper

Section	Analytical goals	Common approaches	Examples
IDA	**Initial data analysis and preprocessing**		
IDA1	Identify inconsistent, suspicious or unexpected values	Visual inspection of univariate and multivariate distributions	Histograms, boxplots, scatterplots, correlograms, heatmaps
IDA2	Describe distributions of variables, and identify missing values and systematic effects due to data acquisition	Descriptive statistics, tabulation, analysis of control values, graphical displays	Measures for location and scale, bivariate measures, RLE plots, MA plots, calibration curve, PCA, Biplot
IDA3	Preprocess the data	Normalization, batch correction	Background correction, baseline correction, centering and scaling, quantile normalization, ComBat, SVA
IDA4	Simplify data and refine/update analysis plan if required	Recoding, variable filtering and exclusion of uninformative variables, construction of new variables, removal of variables or observations due to missing values, imputation	Collapsing categories, variable filtering, discretizing continuous variables, multiple imputation
EDA	**Exploratory data analysis**		
EDA1	Identify interesting data characteristics	Graphical displays, descriptive univariate and multivariate statistics	PCA, Biplot, multidimensional scaling, t-SNE, UMAP, neural networks
EDA2	Gain insight into the data structure	Cluster analysis, prototypical samples	Hierarchical clustering, k-means, PAM, scree plot, silhouette values
TEST	**Identification of informative variables and multiple testing**		
TEST1	Identify variables informative for an outcome	Test statistics, modelling approaches	*t*-test, permutation test, limma, edgeR, DESeq2
TEST2	Perform multiple testing	Multiple tests, control for false discoveries	Bonferroni correction, Holm's procedure, multivariate permutation tests, Benjamini-Hochberg (BH), *q*-values
TEST3	Identify informative groups of variables	Tests for groups of variables	Gene set enrichment analysis, over-representation analysis, global test, topGO
PRED	**Prediction**		
PRED1	Construct prediction models	Variable transformations, variable selection, dimension reduction, statistical modelling, algorithms, integrating multiple sources of information	Log-transform, standardization, superPC, ridge regression, lasso regression, elastic net, boosting, SVM, trees, random forest, neural networks, deep learning
PRED2	Assess performance and validate prediction models	Choice of performance measures, internal and external validation, identification of influential points	MSE, MAE, ROC curves, AUC, misclassification rate, Brier score, calibration plots, deviance, subsampling, cross-validation, bootstrap, use of external datasets

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

IDA

Initial data analysis (IDA) is an important first step in every data analysis and can be particularly challenging in HDD settings. It focuses on understanding the context in which the data were collected, on data cleaning and on data screening

1. Checking the data for technical artifacts such as batch effects or inconsistent values, which can be especially challenging for HDD.
2. Methods for describing HDD data: RLE plots, heatmaps, MA plots, PCA plots
3. Describing and dealing with missing data
4. Methods for preprocessing and normalization that have been specifically developed for HDD are explained.
5. Variable filtering and exclusion of uninformative variables

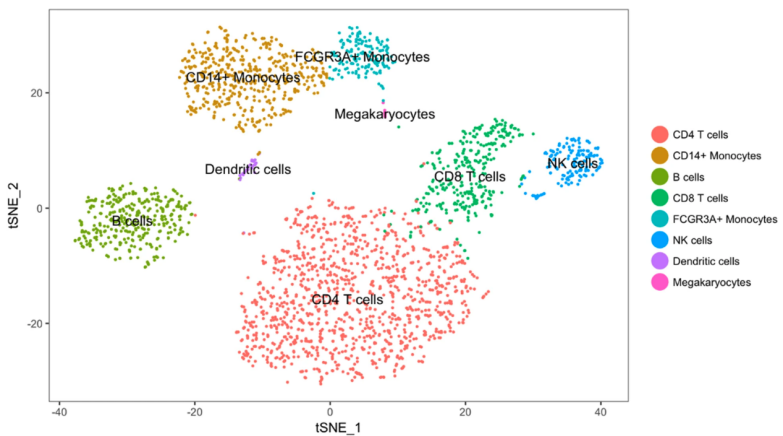
UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

EDA

EDA the goal is to provide an unbiased view of the data

1. to identify interesting data characteristics such as variables with extreme values, associations between variables, or representative subjects with usual values of variables.

From: Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges



Source: "Seurat - Guided Clustering Tutorial". https://satijalab.org/seurat/archive/v1.4/pbmc3k_tutorial.html. citet 2023 Mar 25

2. to gain insight into the structure of the data: Cluster Analysis.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

TEST: Identification of informative variables and multiple testing

Reviews methods for identifying variables informative for an outcome or phenotype, for performing multiple testing, and for identifying informative groups of variables.

1. Test statistics: Hypothesis testing for a single variable (t-tests and permutation tests)
2. Modelling approaches: Hypothesis testing for multiple variables (Limma, EdgeR, DEseq2)
3. Control for false discoveries: Classical multiple testing corrections (Bonferronil, Holm's procedure, Westfall-Young permutation procedure)
4. Control for false discoveries: Methods motivated by HDD (Benjamini-Hochberg, q-values)
5. Sample size considerations
6. Methods for multiple testing for groups of variables (GSEA, Over-representation analysis, Global test, TopGO)

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

PRED: Prediction

A popular goal in biomedical research is the construction of prediction models, but use of extremely large numbers of predictors to build these models presents multiple challenges and risks, if done inappropriately.

1. Construct prediction models (Sparsity, overfitting,)
2. Variable selection
3. Dimension reduction
4. Methods for statistical modelling with constraints on regression coefficients: Ridge regression, lasso regression, elastic net, boosting
5. Methods for statistical modelling with machine learning algorithms: Support vector machine, trees, random forests, neural networks and deep learning
6. Integrating multiple sources of information
7. Assess performance and validate prediction models (Calibration, discrimination, stability, interpretability and practical usefulness)
8. Subsampling, cross-validation, bootstrapping, use of external datasets

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

introduction

We now decided to work in parallel. Currently on three projects,

sample size calculation in HDD:

Co-chairs: Federico Ambrogi, Lisa McShane;
Participants: Harald Binder, Kevin Dobbin, Stefan Michiels, Eugenia Migliavacca, Willi Sauerbrei, Martin Treppner;

influence and choice of the tuning parameters:

Co-chairs: Riccardo De Bin, Lara Lusa, Stefan Michiels;
Participants: Roman Hornung, TBA

use of plasmode data for simulations in HDD:

Co-chairs: Axel Benner, Jörg Rahnenführer;
Participants: TBA

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

sample size calculation in HDD

A research protocol should specify the study design, including planned sample size

depends on the primary endpoint, analysis goal, and other key assumptions.

In HDD settings, traditional sample size calculations break down due to:

the large number of hypotheses tested;
complex modeling or analysis strategies typically employed;

Several approaches for sample size calculation tailored to certain HDD settings have been proposed in the literature:

utility and uptake has not been systematically evaluated;
it is unclear what kind of sample size justification is used;
if any justification is used at all.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

sample size calculation in HDD

Currently screening the literature. Two reviews underway:

methodological review,

1. identify statistical methods papers describing HDD samples size methods;
2. starting from specific "key words" for "study design criterion" and "Study goal or method" / "Data type";
3. open to augment the list during the applied literature search;
4. record number of literature citations for each identified method.

Study goal or method	Study design criterion	Data type
Artificial neural network	Brier score	Big data
Artificial intelligence	Calculating number of . . .	Carbohydrate array
Boosting	Calculation of number . . .	CGH array
Classification	Cases needed . . .	Comparative genomic hybridization/CGH
Classification and Regression Trees (CART)	Classification accuracy	ctDNA array
Classifier	Design of study(ies)	Deep Sequencing
Classifier development	Designing . . . study(ies)	DNA array
Cluster discovery/discover clusters/discovering clusters	Determination of number . . .	DNA sequencing
Deep learning	Determining number . . .	Electronic health record (EHR)
DESeq	Discrimination accuracy/ability	Epigenomic(s)
DESeq2	Estimating number of . . .	Epigenomic(s)
Discriminant analysis	Estimation of number . . .	Exposome/exposomic
edgeR	False discovery(ies)/False discovery(ies) rate	Gene expression microarray(s)
Elastic net	FDR	Gene expression profiling/profiles
Find/finding clusters	Individuals needed . . .	Gene panel data
Find/finding latent class(es)	Misclassification rate	Gene sequencing
Find/finding structure	Number needed . . .	GeneChip
Find/finding subtypes	Number of cases	Genome-wide association study (GWAS)
Graphical models	Number of individuals	Genomic data
Identify cluster(s)/identifying clusters/cluster identification	Number of participants	Genomic studies
Identify latent class (s)/identifying latent classes/latent class identification	Number of replicates . . .	Genomic(s)
Identify structure(s)/identifying structure/structure identification	Number of samples	Glycan array

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

sample size calculation in HDD

applied review,

1. identify which methods (if any) are actually being used in applied/biomedical papers.
2. papers need to deal with HDD (data of specific types);
3. focus on the 15 journals with the highest impact factor in selected fields (i.e. oncology) in a defined time interval;
4. extended to top 5 (impact factor) for general medicine journals, and possibly biology/biomedical ones . . .
5. compute the percentage of studies (within HDD), where a sample size calculation was performed;
6. record the method and/or the justification used;

   it involves searching for sample size justification in the text;
7. scope might be further restricted if the number of publications identified will result too large.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

sample size calculation in HDD

Example of citation analysis of methods papers:

1. Start from the collected methodological literature
2. Retrieve citations of methodological papers
3. Distinguish among citations from applied, guidelines and other methodological papers.

METHODS	CITATIONS	Applied	Guidelines/Reviews	Methods	Other	software
Power and sample size estimation in high dimensional biology	59	8	29	20	2	
Sample size planning for survival prediction with focus on high-dimensional data	2		2			
Sample size determination for high dimensional parameter estimation with application to biomarker identification	0					
Sample Size and Power Calculation for Molecular Biology Studies	10	8	1	1		
FDR-controlling testing procedures and sample size determination for microarrays	41	3	12	26		
Power and sample size calculations for high-throughput sequencing-based experiments	26	9	11	5		
Size matters: how sample size affects the reproducibility and specificity of gene set analysis	20	7	2	11		
Harmonization of quality metrics and power calculation in multi-omic studies	23	6	14	3		
A simple method for assessing sample sizes in microarray experiments	79	29	23	27		
General power and sample size calculations for high-dimensional genomic data	17	6	1	10		
The effects of sample size on omics study: from the perspective of robustness and diagnostic accuracy	1		1			
RnaSeqSampleSize: real data based sample size estimation for RNA sequencing	39	31	3	5		
False discovery rate, sensitivity and sample size for microarray studies	345					
Feasibility of sample size calculation for RNA-seq studies	18	5	5	8		
Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array	124	97	16	10		

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

sample size calculation in HDD

Goals:

describe the methodologies most used in applied research, distinguishing between:

- Class discovery;
- Class prediction;
- Class comparison;

provide examples when software is available;

provide recommendations for applied researchers;

have a starting point for future methodologic work.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

influence and choice of the tuning parameters

Statistical and Machine Learning models for classification or prediction, especially in HDD, strongly rely on tuning parameters,

- choosing the best value of the tuning parameter is often more important than choosing the method;
- Especially for HDD, this tuning process can result in overoptimistic prediction performance estimates.
- often obtained data-driven.

Unfortunately, in many cases:

- there is no understanding on the role of the tuning parameter;
- default values do not work in broad generality;
- especially when derived in low dimensional contexts.

Other issues:

- when data-driven, tuning parameters must be computed on a dedicated subset of the data (validation set $\neq$ test set);
- complex methods with many tuning parameters are very hard to handle.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

influence and choice of the tuning parameters

Goals:

describe the role and the importance of the tuning parameters;

describe typical procedures used to find them;

provide examples, using real high-dimensional data

chronic obstructive pulmonary disease dataset;

discuss common issues related to the tuning parameters;

provide recommendations for their choice in practice.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

use of plasmode data for simulations in HDD

Simulation studies are especially challenging for HDD, yet they are essential tools needed to perform evaluation and comparison of different methods.

The typical approach for simulation studies is to use synthetic data, for which the entire true data generating process is known ("parametric" simulation).

Parametric simulations of HDD are usually based on overly simplistic assumptions about the high-dimensional multivariate data distribution.

A plasmode dataset preserve a realistic data structure by resampling covariate data from real-life datasets

and using parametric model to generate outcome;

This idea seems particularly promising for HDD, but its usefulness has not yet been properly evaluated in the literature.

Goal:

Evaluate potential of using plasmode datasets in high dimensional simulation studies.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Current work:

use of plasmode data for simulations in HDD

Ongonig research projects (preparations for STRATOS project):

- comparison of plasmode approaches,
  - A. Benner, N. Schreck (DFKZ, Heidelberg), A. Slynko (University of Waterloo), M. Saadati (Statistical Consulting);
- comparison of methods for quantifying dataset similarity.
  - J. Rahnenführer, M. Stolte, A. Bommert (TU, Dortmund);
- comparison of parametric and plasmode approaches for simulation studies in LDD,
  - all researchers mentioned above.

STRATOS TG9 project (starting in summer 2023)

- comparison of parametric and plasmode approaches for simulation studies in HDD,
- how far must the parametric assumptions deviate from the true model for the plasmode approach to be superior?

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

Conclusion

Proliferation of high-dimensional data in biomedical research has brought unprecedented opportunities to advance knowledge. In order to harness the power of the rapidly evolving repertoire of analysis methods to reveal useful insights from HDD, it is imperative that researchers have access to guidance on the methods available and their proper application

Visit https://www.stratos-initiative.org/group_9

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ